Managing Bandwidth: The Key to Cloud-Assisted Autonomous Driving

背景

普遍认为，对于重要的实时控制系统，例如自动驾驶，不能依赖云。

因为自动驾驶有着严格的实时执行的限制，同时蜂窝网络是高度可变的，且有性能约束。

而作者认为可以，且必须能。

因为随着ML模型大小的增加，硬件的改进和移动网络的发展，作者认为，有机会把时间敏感和延迟关键的计算卸载到云，不过需要仔细分配带宽，以满足严格延迟的SLO。

关于自动驾驶

自动驾驶（AV）系统中，计算通常是按流水线（pipeline）结构组织的，每个组件负责执行一个特定的任务，例如物体检测等，这些任务大多基于机器学习（ML）模型。

但是，**ML模型的准确率越高，计算越密集，延迟也越高。**因此，构建一个自动驾驶流水线时，需要仔细考虑**精度与响应时间之间的权衡。**

为什么适合把每个组件当成一个服务?
自动驾驶系统的流水线结构非常适合将每个组件视作一个"服务"，原因有几点：

1. **每个组件的任务明确**：每个组件都有明确的目标和功能，比如物体检测、路径规划等。每个组件的功能类似于一个独立的服务，它接收输入、执行任务并产生输出。
2. **组件的接口清晰**：这些组件的任务定义了一个具体的接口（例如：输入是什么，输出是什么），这使得它们作为独立的服务可以方便地进行管理和调度。
3. **多种实现选择**：每个组件的实现可以从不同的机器学习模型和算法中选择，且不同的模型或算法会有不同的延迟和精度特性。通过将每个组件作为一个服务，可以**灵活选择最适合的实现方式**，根据实际情况调整模型，以平衡精度和响应时间。
4. **服务的独立性和可扩展性**：把组件当成服务可以提高系统的模块化，使得每个组件可以独立开发、测试和更新，而不影响整个系统的运行。同时，如果需要更高的性能或更低的延迟，系统也可以灵活地替换掉某个组件的实现。

AV流水线中基于ML的关键组件

- 传感器：每秒钟产生超过8Gb的数据，并提供AV周围环境的快照
- 感知模块：对传感器数据流进行处理，融合到以自身为中心的地图中，标注附近的障碍物、可行驶区域和交通法规。感知模块执行几种不同的任务，如目标检测、目标跟踪和车道线检测，它们使用不同的模型并形成**子服务。**
- 预测模块：预测附近物体的动作。通过处理感知模块的输出，利用Transforms等计算密集型神经网络，基于物体的动作和行为进行预测。
- 规划模块：生成AV的安全舒适的移动方案。传统的规划器一般采用搜索算法，而未来的规划器可能采用ML的方式。

一些Q&A

为什么即使现在有更先进的（SOTA，State-of-the-Art）模型，AV制造商仍然不使用它们，使发布的自动驾驶车辆安全性可能低于预期呢?

SOTA模型的确能提供更高的性能和安全性，但目前的自动驾驶车辆已经足够安全，可以进行公开测试。

尽管SOTA模型在理论上可以进一步提高安全性和性能，但当前车辆的计算能力可能不足以支持这些复杂的模型。

为什么不在车上安装更好的硬件，而选择使用云计算资源来支持自动驾驶功能?

1. **功耗和散热限制**：虽然更强大的硬件（如SOTA GPU）是现有的，但它们通常需要高功率和良好的散热环境，这对于车载硬件来说是一个挑战。汽车对功耗和散热有严格的要求，而现代GPU（如NVIDIA H100）通常功耗非常高，并且产生大量热量，这使得它们不适合直接部署在汽车上。
2. **经济性问题**：新的计算硬件非常昂贵。例如，2024年的NVIDIA H100 GPU价格为40,000美元，而一辆新特斯拉的价格约为30,000美元。即便是自动驾驶硬件的成本也比这些SOTA云端硬件便宜10倍左右。
3. **统计复用的机会**：汽车通常每天只使用约60分钟，剩下的23小时大部分时间是空闲的。这为云计算资源的共享提供了机会。通过在云端共享计算资源，很多车辆可以共同使用一个计算池，从而提高资源利用率。并且，云计算资源可以在负载较低时与非自动驾驶的工作负载共享计算能力。
4. **云计算不排除改进车载计算**：尽管云计算可以提供强大的支持，但这并不意味着车载计算就不能改进。如果未来有更高效的硬件（如更低功耗和更小体积的硬件）出现，自动驾驶厂商仍然可以将其部署到车上。
5. **灵活性和选择**：文中提出的方案是增加灵活性，允许自动驾驶厂商根据不同需求和技术进展选择何时投资车载计算硬件或远程计算资源。厂商可以根据实际情况决定采用更强大的车载硬件，或继续利用云端资源。

硬件的开销是不是快速减少的?

是，但是模型大小是快速增加的，并且汽车寿命往往在10年以上，那些老旧汽车需要一些方式来受益于更大模型的新能力。

既然硬件成本低，为啥不升级旧汽车的硬件?

自动驾驶汽车的公司可以在拥有、运营和维护所有车辆时进行升级。然而，私人车辆的升级并不那么容易。尽管召回对消费者来说是免费的，但目前的修复率是 52-64% 。

为什么不将模型压缩得更小，以便能够在部署的硬件上运行?

模型压缩通常会降低精度，同时压缩模型仍然可能太大或太慢而无法在 AV 上运行。

怎么能依赖云端计算？如果网络不可靠，或者云因其他原因出现故障怎么办?

1. **云端计算的角色**：回答强调，云端计算主要是用来“贪心地”提升性能，而不是基础功能的必需品。也就是说，云端计算的目的是在车辆的基本安全性和功能（例如当前车载计算的流水线）已经达到并超越人类驾驶员的安全极限时，通过云端进一步增强车辆性能。
2. **车载计算已经足够安全**：现有的车载自动驾驶系统已经在安全性上超越了人类驾驶员，满足了基本的安全要求。这意味着，**即使云计算资源不可用（例如网络连接失败），车载系统仍然能够独立运行，保证安全性。**
3. **云端是性能的增强**：云端计算的加入是为了进一步提升车辆的性能，比如使用更强大的计算资源进行复杂的模型推理或数据处理，但它并不是自动驾驶系统的核心或必需部分。即使云计算出现故障，车辆仍然能够依靠本地的计算能力继续安全行驶。

可行性分析

性能

基于先前的5G网络测量，假设有一个连接到附近的数据中心的 5G 连接，RTT 为 12 毫秒，上传带宽为 200 Mbps。

对于 EfficientDet（ED) 对象检测模型的变体ED0-7

Model			Compute		Network	Total	
Name	Accuracy	Input Size	Orin [ms]	H100 [ms]	Transfer [ms]	H100 + Transfer [ms]	Speedup
ED0	34.3	512 × 512	112	26	20	46	2.4×
ED1	40.2	640 × 640	136	32	24	56	2.4×
ED3	47.2	896 × 896	325	41	36	77	4.2×
ED5	51.2	1280 × 1280	1067	65	52	117	9.1×
ED7	53.4	1536 × 1536	1955	101	83	184	10.6×

发现，推理时间的改进，超过了网络延迟。

对于推理速度，H100 执行 ED 模型的速度比 Jetson Orin 快 4 – 19 倍，使得以前不可行的模型能够在 SLO 内运行，即使考虑到网络时间。

经济

网络成本

商业网络使用费主要按GB收取。

Rank	Country	$/GB	$/Hour
1	Singapore	$0.07	$1.65
2	Netherlands	$0.36	$8.04
3	Norway	$2.09	$47.07
4	United States	$0.75	$16.88
5	Finland	$0.26	$5.81
–	China	$0.27	$6.14
–	Israel	$0.001	$0.02
–	*10th pct*	$0.062	$1.39
–	*Median*	$0.37	$8.42

在蜂窝数据价格较高的国家，运营商可能会选择通过有选择地利用远程资源来降低成本。

但是，蜂窝数据成本是呈强劲下降趋势的，从 2019 年到 2024 年，全球每 GB 的中位价格下降了 4 倍，从 5.25 美元降至 1.28 美元。

云计算成本

云服务提供商（如Lambda Labs）提供按小时计费的GPU租赁服务，价格从
$0.80（NVIDIA A6000）$到$2.49（NVIDIA H100）$不等

通过共享计算资源（例如多辆车共同使用云端GPU），可以显著降低每辆车的计算成本。

车辆使用率：

- **个人车辆**：美国平均每位司机每天驾驶约60.2分钟，即车辆的利用率仅为4.2%。大多数时间，车辆是空闲的，无法充分利用计算资源。
- **自动驾驶网约车**：相较于个人车辆，自动驾驶网约车（如Uber）的利用率较高，预计约为59%。这些车的使用率相对较高，可以更好地利用云计算资源。

云计算的成本效益：

- 如果以购买一台H100 GPU（约$40,000）的成本为基准，对于普通的个人驾驶车辆（4.2%的使用率），这种GPU的购买成本相当于需要44年的云端租赁费用。而对于自动驾驶网约车（59%的使用率），则相当于3年。
- 云计算资源提供了灵活性，车主和运营商可以根据计算需求和成本敏感度选择使用不同的计算资源（如GPU）。此外，远程资源无法在事故中被损坏或盗窃。
- 通过优化资源利用率（例如批处理和请求调度），云端计算还可以进一步提高效率，降低成本。

但从经济角度来看，使用云计算而不是购买昂贵的车载专用硬件（如H100 GPU）是更具成本效益的选择。

方法

方法的目标是选择组成 AV 控制流水线的服务子集在云中运行，以及如何在它们之间分配带宽。

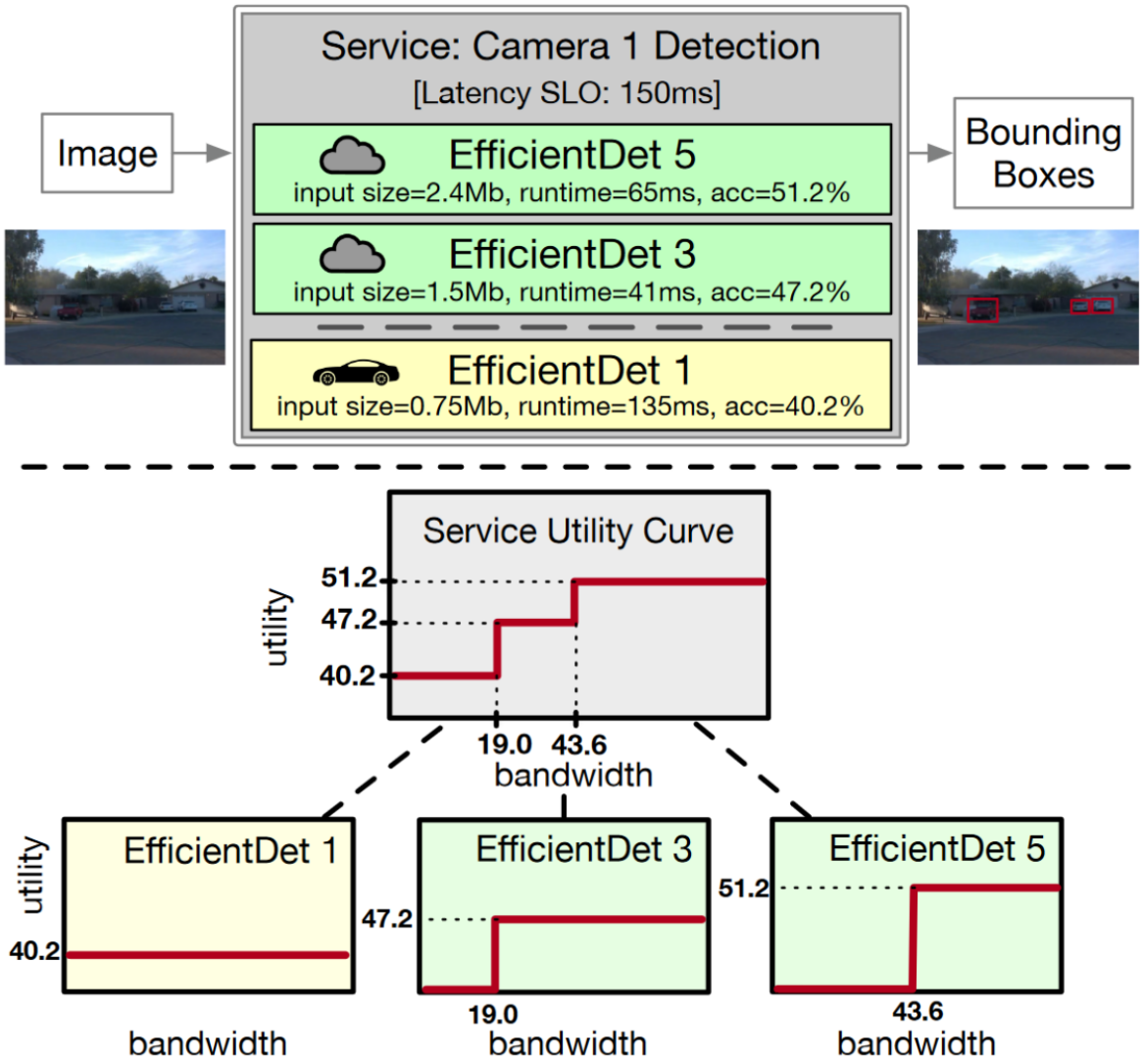
这里**使用感知模型，如ED系列的模型作为例子来说明。**

三个关键：

- 为 AV 流水线中的每个服务设置延迟 SLO。服务能否满足SLO，依赖于模型的运行时间和数据的传输时间。
- 带宽决定数据传输的速度。给一个服务分配更高的带宽，能让它更好的满足SLO。由于下行带宽是上行带宽的10倍以上，且目标检测的输出（感知模型输出）相对较小，时延=传播时间（ping RTT）+上行数据传输时间。
- 准确性衡量 ML 模型执行任务的情况。例如，目标检测使用mAP（平均精度）作为准确性指标。本文引入了先前工作的效用曲线（表示分配的带宽如何为应用程序提供效用），并用**服务准确性代表效用。**

对于物体检测服务，车上使用ED1，云上使用ED3和5。

ED1 是在SLO内车上能运行的最准确的模型，ED3和5提供了更高的精度，但为了满足SLO，只能在云上的的硬件运行。



若服务的可用带宽太低，将会违背SLO，导致效用为0。

组合曲线提供了我们服务所能达到的给定带宽的最大效用(即,准确性)，而车载模型提供了一个保证的最小值。

收获

- 实际的驾驶环境是动态的，需要动态的效用曲线。随着环境的变化，如果观测数据与训练数据不同，可能会导致模型精度下降，需要AVs监控性能并更新带宽分配。同时，针对特定的环境可以采用特定的模型。
- 带宽不光可以空分复用，还可以**时分复用**。若按时分复用，服务的时隙中，可以使用**所用可用带宽**传输数据。这是，我们需要优化的就是服务的数据传输顺序，来最小化延迟，这类似于先前的调度工作。
- 大规模的部署，会增加资源竞争，但也会产生新的机会。当越来越多的AV依赖远程资源，蜂窝网络的负载会进一步增大，特别是上下班高峰期。此外，随着规模增大，云上的资源可用性也会产生问题，怎么让云成为无限资源。但是随着规模增大，**AVs之间也可以相互合作，例如共享数据以减小盲区**。
- 蜂窝网络的带宽是延迟的主要来源。其中，设备到单元基站的带宽是一个瓶颈。将计算资源部署在靠近基站的地方并不会显著降低延迟，因为设备到基站的带宽才是主要的限制因素，所以把云部署在**附近的数据中心**已经足够满足低延迟的需求。

- 远程资源具备成本效益。虽然远程H100 GPU的计算价格较高，约为每小时2.49美元，而部署车载H100时的摊销成本为每小时2.28美元，但是，**车辆的低利用率使得远程资源更具成本效益**，因为闲置的远程资源可以被重新利用。